Relevance Segmentation of Long Documents

Zsolt Szántó, Alex Sliz-Nagy, István Nagy T., Ádám Csuma-Kovács,
Veronika Vincze, Richárd Farkas

Black Swan Hungary,
Szeged, Tisza Lajos krt. 47.,
szeged@blackswan.com

Abstract: In this paper, we present our methods to identify the most salient topics for a selected domain based on topic modeling. We propose a topic relevance score and segmentation procedure which can split the document into parts referring to various topics. We also offer a solution for visualizing textual spans that are related to a given topic. In this way, it can be easily determined which are the most relevant and most irrelevant segments of a long document (like blog posts or news articles).

1 Introduction

Nowadays, a huge amount of textual data is published every day on the internet, in the form of weblogs, social media posts, posts on official websites etc. However, the large amount of data makes it impossible to process it manually -- for instance, when the user is interested in a topic, the number of documents related to the specific topic might be overwhelming and thus, no human can easily find all the relevant documents. The problem might be even more difficult, considering that a single document can contain several topics itself, some (or all) of which might be relevant to the user. If only a smaller segment of a long document, like a blog post or news article, is relevant to the user, he should not waste time with reading the whole document.

There are various solutions for reducing the human processing time of a long document, for instance, keyphrase extraction (which assigns a number of short phrases to documents, representing the content) [1,4] or document summarization (which offers a few sentence long summary of the whole document) [6].

In this paper, we propose an automatic document segmentation and relevance visualization tool designed for long documents. Our solution is built on the top of LDA topic modelling. The information target of the user is defined through a set of keywords. We first rank the LDA topics by relevance to the input of the user, then segment a long document by assigning a topic for each word smoothed in word sequencing. We also offer a solution for visualizing textual spans that are related to a given topic. In this way, it can be easily detected how many topics occur in a document, which topics are the most salient ones and which are more marginal. We demonstrate our system on Hungarian news articles.

2 Literature review

Topic modeling aims at discovering the abstract topics that occur in a document or set of documents. In other words, it discovers the hidden semantic structures of a text: certain words are expected to occur in connection with certain topics, and their presence strongly indicates that the document is about that given topic. Documents usually consist of several topics, and the most salient topics can easily be identified with topic modeling.

Latent semantic indexing (LSI) is a widely used method to transform the original document vectors to small-dimension vectors [3]. In probabilistic LSI, each word in a document can be seen as a pattern of a mixed model, composed of different topics. Thus, a document can be seen as a mix of different topics.

Another model for representing topics is Latent Dirichlet Allocation (LDA) [2]. It is a generative statistical model allowing observations to be explained by unobserved groups for data similarity. Each document is seen as a mixture of a small number of topics and each word is attributable to one of the document's topics. LDA is an example of a topic model and was first presented as a graphical model for topic discovery.

Topic segmentation has also been widely investigated. An early attempt to identify topic boundaries in free texts was reported in [7]. A new hierarchical Bayesian model was proposed for unsupervised topic segmentation [5] and LDA is also frequently used for text segmentation [8].

3 Methodology

For our experiments, we downloaded approximately 100,000 documents from different Hungarian news portals (e.g. www.index.hu, www.origo.hu) then tokenized and lemmatised them with magyarlanc [9].

Our method requires two databases, the first one is a set of documents and the second is a manually created set of words that describe a topic. We would like to calculate similarity between this predefined topic and the documents.

3.1 Preprocessing

We used standard preprocessing methods, we removed the stopwords, punctuations, and numbers and we lemmatized [9] the corpora.

3.2 Ranking of the documents

First we ranked the topics based on the similarity between a topic and our predefined topic indicator word list. We ran LDA on the lemmatized dataset and we calculated r_t ranking score to each topic t, with the following equation:

$$r_t = \sum_{\substack{w<-indicator \\ words}} p(w|t)$$

The topics with highest r_t are more similar to our predefined word list.

Based on this topic ranking scores we can get a ranking over the documents. The r_d document ranking score for document d is derived from this equation:

$$r_d = \sum_{t<-topics} p(d|t) * r_t$$

Now we have a ranking over the documents where the top of this list is highly related to the predefined topic.

3.3 Topic detection inside the document

Some of the documents contain more than one topic, so we developed a method which can extract these topics inside a document. We used Hidden Markov Model (HMM) to separate different topics in one document. In our HMM the words are the observations and the topics are the hidden states, each observation are generated from a hidden state (like in POS tagging). The HMM requires two parameter matrices as input, the emission probability matrix that describe a distribution of the observations over the hidden states and the transition probability matrix that describes the probability of a transition between two states.

The LDA calculates the distribution of words over a topic, which we can directly use as the emission probabilities in HMM.

For transition probability we only use two values, one for the situation when we keep the previous state, one when we change the state:

$$p(t_i, t_j) = \begin{cases} \alpha, & i = j \\ \dfrac{1-\alpha}{|T|-1}, & otherwise \end{cases}$$

where α is a parameter that determines the size of the topics.

By using the Viterbi algorithm, we can get the topic sequence over the words of document d with the highest probability.

4 Results

In our experiments, we focused on two domains, namely, sports and music. We used Wikipedia-based lists to construct an initial seed list for describing the domains. Here we present our results for ranking topics in connection with the domains and we also report on how documents can be segmented on the basis of topics present in the document.

id	top words					score
39	magyar Hungarian nemzetközi international	olimpiai olympic méter meter	verseny competition nyer win	olimpia Olympics két two	szövetség association hosszú Hosszú	5.2501
9	magyar Hungarian két two	nyer win női female	első first második second	hely place döntő final	csapat team férfi male	4.1828
6	hely place kör lap	első first futam race	autó car tud can	verseny race második second	két two motor motor	4.1565
0	meccs match első first	csapat team játékos player	két two szezon season	pont point nyer win	ben in Nfl NFL	4.0798
47	oldal site twitter Twitter	tud can ember man	facebook facebook szeptember September	fotó photo 24hu 24.hu	címlap title page kép picture	4.0561
2	csapat team mérkőzés match	meccs match gól goal	válogatott national team bajnokság league	játékos player klub club	játszik play pálya pitch	4.0494
28	terület area épít construct	épület building projekt project	terv plan épül build	gép machine beruházás investment	város city Park Park	4.0454
22	manos Manos annyi many	film film látható visible	nap day munka work	szereplő actor néz watch	sztori story Lát See	4.0060
34	magyar Hungarian lászló László	egyetem university györgy György	ben in budapesti Budapest	budapest Budapest alapítvány foundation	istván István Program Program	4.0018
35	kutató researcher talál find	kutatás research állat animal	föld earth tanulmány study	egyetem universiry tudományos academic	víz water eredmény result	4.0018

Table 1: Salient topics related to sports.

id	top words					score
46	dal	szám	zene	zenekar	koncert	10.929
	song	hit	music	band	concert	
	fesztivál	videó	lemez	című	album	
	festival	video	record	entitled	album	
23	kis	hely	például	szép	inkább	8.7910
	small	place	for example	nice	rather	
	étterem	kicsi	egészen	két	név	
	restaurant	small	totally	two	name	
40	fotó	kép	épület	magyar	Hungarian	8.1943
	photo	picture	building	Hungarian	Budapest	
	tér	múzeum	jános	lászló	józsef	
	square	musem	János	László	József	
15	ben	kap	díj	első	két	8.0201
	in	get	prize	first	two	
	három	idén	vesz	név	nyer	
	three	this year	buy	name	win	
10	film	című	rendező	sorozat	színész	7.9871
	film	entitled	director	series	actor	
	mozi	történet	rész	jelenet	néző	
	movie	story	part	scene	audience	
6	autó	hely	első	tud	két	7.9504
	car	place	first	can	two	
	kör	második	futam	motor	hamilton	
	lap	second	race	motor	Hamilton	
37	európai	magyarország	magyar	ország	kormány	7.9076
	European	Hungary	Hungarian	country	government	
	bizottság	uniós	unió	európa	orbán	
	committee	Union	Union	Europe	Orbán	
32	manos	film	nap	szereplő	sztori	7.9054
	Manos	film	day	actor	story	
	annyi	látható	munka	néz	Lát	
	many	visible	work	watch	See	
42	gép	föld	hajó	két	kutató	7.8989
	machine	earth	ship	two	researcher	
	első	tud	repülőgép	nap	tudós	
	first	can	airplane	day	scientist	
19	facebook	oldal	rendszer	cég	használ	7.8964
	Facebook	site	system	company	use	
	tud	telefon	eszköz	felhasználó	google	
	can	telephone	device	user	Google	

Table 2: Salient topics related to music.

Tessék, lányok, akik nem szültek, mégis rengeteg gyereket tettek újra nagyon boldoggá
Megmentették a három új-zélandi tehenet, akiket a földrengés vágott el a világtól
Végre egy budapesti építkezés, ahol az épülethálón nem reklám van, hanem a leendő homlokzat
Ha Németh Szilárd cikket olvas, a brazíliai őserdőkben felsír egy fa
Mindent az olvasóért: megkóstoltam a Nutellaburgert, hogy önöknek már ne kelljen
Rogán azt állítja, kávézni ugrott fel régi ismerőséhez. Vajon ezúttal sikerült igazat mondania?
Tessék, lányok, akik nem szültek, mégis rengeteg gyereket tettek újra nagyon boldoggá
Emlékszik még a Balatont hosszában átúszó lányokra?
Nem csak úszásban és példamutatásban jók, de most a Kispingvin Úszó és Vízilabda Előkészítőben úszó gyerekekkel közösen gondoltak egyet és nekiálltak rászoruló gyerekeknek ajándékot gyűjteni. Ez olyan jól sikerült, hogy a Magyar Máltai Szeretetszolgálat Családok Átmeneti Otthonában élő mind a 87 gyermeknek jutott egy hatalmas névre szóló doboz! A dobozokat a Máltai Szeretetszolgálat Jézuskája karácsonykor a megfelelő fa alá fogja tenni.
Ez a világ legnagyobb magányosan álló jármű-monstruma
Németország keleti részében található meg a mai napig ez a hatalmas, 112 métere hosszú, 57 méter magas szerkezet. Melyet a kelet-német ipar egykori dicsőjét mutatja. Ez a hatalmas kotrógép jelenleg a világ legnagyobb magára hagyott járműve. A gép elején látszódnak hatalmas "fogai" vagyis vödrei, amivel szó szerint ette a földet. Ezekkel darabonként 15 köbméter földet tudtak kibányászni egy kör alatt anno.

Fig. 1. Sample text for visualizing topics related to sports.

4.1 Ranking topics for domains

The ten most salient topics related to sports are presented in Table 1. As can be seen, the first 6 topics contain lots of sports words. There seems to be one outlier topic, which is ranked 5th: topic 47 is related to social media (including words like Facebook and Twitter). However, sports events are often advertised and reported in social media, hence the frequency of social media vocabulary can be easily explained in the sports domain as well. Also, topic 28 describes construction works, which again might be connected to sports, for instance, when constructing buildings for sports facilities such as stadiums, sports halls or football pitches.

The ten most salient topics for music are presented in Table 2. The first topic is unambiguously related to music. There is a huge gap between the scores for the first and the second most salient topic, which suggests that the vocabulary of music is utterly distinct from all the other topics. However, topics 15 and 10 might be also loosely related to music, as there are music awards where prizes can be won (topic 15) and films are also accompanied with music (topic 10). Moreover, topic 0 may be of relevance as well: there are several recent reports on sexual harassment from the entertainment industry, so unfortunately a topic on sexual abuse can also be connected to the music domain.

4.2 Document segmentation

Here we illustrate our results on segmenting the documents on the basis of the topics mentioned. For this purpose, we made use of the *Mindeközben* (Meanwhile) column of the news portal *index.hu*, which includes short pieces of news of miscellaneous topics, hence they are supposed to contain multiple topics.

Figure 1 shows a sample from a document where different topics are marked with different colors. Text spans which are related to sport and have a high position in the topic ranking are highlighted with green (and with bold font).

Figure 2 shows another sample from the *Mindeközben* column. Here, green spans denote textual content related to music. As can be seen, the first sentence of the document contains an invitation to a music festival, which is then followed by titles of other short news. Later, the music festival is described in full detail, which was also identified as belonging to the music topic by the algorithm.

Kíváncsi, hogy mulattak a francia udvarban? Menjen a Régizene Fesztiválra, ahol olyat hallhat, amit 274 éve senki

Megmentették a három új-zélandi tehenet, akiket a földrengés vágott el a világtól

Végre egy budapesti építkezés, ahol az épülethálón nem reklám van, hanem a leendő homlokzat

Ha Németh Szilárd cikket olvas, a brazíliai őserdőkben felsír egy fa

Mindent az olvasóért: megkóstoltam a Nutellaburgert, hogy önöknek már ne kelljen

Rogán azt állítja, kávézni ugrott fel régi ismerőséhez. Vajon ezúttal sikerült igazat mondania?

Március 2-án kezdődik a Müpában 2. Régizene Fesztivál, s mivel már az előzőért is nagyon odavoltunk , nem mehetünk el szó nélkül a mostani mellett sem. Bár lesznek azért alapdarabok is a műsoron (a Capella Savaria például Bach hat Brandenburgi versenyét adja elő), a központi téma a Magyarországon nem annyira ismert francia barokk lesz.

A legnagyobb durranásnak Mondonville Isbé című operájának koncertszerű előadása ígérkezik vasárnap az Orfeo Zenekar és a Purcell Kórus közreműködésével, Vashegyi György vezényletével.

Ezt az operát ugyanis keletkezése, 1742 óta még soha nem tűzték műsorra!

Akit jobban érdekel a téma, az már egy nappal korábban, szombaton odaballaghat, mert két ingyenes előadás is lesz, az egyiket Vashegyi, a másikat Benoît Dratwicki (a Versailles-i Barokk Zenei Központ művészeti vezetője) tartja a Mondonville-ről és a francia operáról.

A szívemhez legközelebb álló előadás Sebestyén Mártáéké lesz. Ő Andrejszki Judittal, egy valódi provanszál trubadúrral, Miquèu Montanaróval, s annak a fiával Baltazárral lépnek a színpadra. Ahogy ígérik, az "autentikus francia barokk zene mellett a francia népi hangzásvilág és a magyar népzene találkozásának lehetünk fültanúi."

Ilyen az, amikor 20 ezer kacsa egyszerre kel át az úton

Az eset egyébként Kínában történt, és ránézésre óriási dugót okozott, de hát a kacsák szabályosan közlekedtek, így az autósoknak egy szavuk sem lehet.

Benjamin Franklin, az Egyesült Államok Függetlenségi nyilatkozatának egyik aláírója 1790 óta egy philadelphiai temetőben nyugszik, vagyis próbál nyugodni, de a temető látogatói folyton pénzérméket dobálnak a sírjára, évente több tízezer pennyt, mivel Franklin egyszer azt mondta: „Ha félretettél egy pennyt, kerestél egy pennyt” - írja az MTI.

Meg is lett a baj: az időjárás viszontagságainak kitett márványtábla elrepedt, a folytonos baszkurálástól egyre nagyobb lett a repedés, most meg nincs pénz helyrehozni, pedig a sírt kezelő alapítvány még támogatást is kapott rá. De még így is hiányzik tízezer dollár (nagyjából hárommillió forint), amit adománygyűjtéssel szeretnének összehozni. Pedig lehet, hogy csak meg kéne kérni a turistákat, mostantól inkább egydolláros bankókat hajigáljanak, az legalább nem töri össze a sírt, és hamar össze is jönne az összeg, szóval mindenki jól járna, és Franklin is foroghatna tovább a sírjában.

Fánkcsapdával és tökéletes öniróniával figyelmeztet az augustai rendőrség

Ugye ön is ismeri azt a sztereotípiát, hogy az összes amerikai rendőr dagadt és szeret fánkot zabálni? Nos, az augustai rendőrök nemcsak ismerik, de szívesen építenek is rá, ha a lakosságot kell figyelmeztetniük valamire, jelen esetben épp arra, hogy csapdák kihelyezéséhez engedélyre van szükség.

Mindezt egy zseniális Facebook bejegyzésben adták az emberek tudtára: egy rendőr egy doboz ingyen fánk mellett guggol, alá pedig azt írták, hogy bárki is teszi ki ezeket a fánkcsapdákat, most már igazán abbahagyhatná, mert Chase már harmadszor késik el miattuk a héten.

Fig. 2. Sample text for visualizing topics related to music

5 Conclusions

In this paper, we presented our methods to identify the most salient topics for a selected domain in Hungarian news articles, based on topic modeling. We also showed a solution for visualizing textual spans that are related to a given topic, focusing on the sports and music domains. In this way, it can be easily detected how many topics occur in a document, which topics are the most salient ones and which are more marginal with regard to the central topic of the document.

References

1. Berend, G., Farkas, R.: Keyphrase-Driven Document Visualization Tool. In: The Companion Volume of the Proceedings of IJCNLP 2013: System Demonstrations (2013) 17-20
2. Blei, D., Ng, A., Jordan, M.: Latent Dirichlet Allocation. Journal of Machine Learning Research, Vol. 3, No. 5 (2003) 993-1022
3. Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., Harshman, R.: Indexing by latent semantic analysis. Journal of the American Society for Information Science. Vol. 41, No. 6 (1990) 391-407
4. Kim, S.N., Medelyan, O., Kan, M-Y., Baldwin. T.: SemEval-2010 task 5: Automatic keyphrase extraction from scientific articles. In: *Proceedings of the 5th International Workshop on Semantic Evaluation* (SemEval '10). Association for Computational Linguistics, Stroudsburg, PA, USA (2010) 21-26
5. Lan, D., Buntine, W., Johnson, M: Topic Segmentation with a Structured Topic Model. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Atlanta, Georgia Association for Computational Linguistics (2013) 190-200
6. Qazvinian, V., Radev, D. R., Mohammad, S. M., Dorr, B., Zajic, D., Whidby, M., Moon, T.: Generating extractive summaries of scientific paradigms. *J. Artif. Int. Res., Vol.* 46, No. 1 (2013) 165-201
7. Reynar, J. C.: Statistical models for topic segmentation. In: *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics* (ACL '99). Association for Computational Linguistics, Stroudsburg, PA, USA (1999) 357-364
8. Riedl, M., Biemann, C.: Text Segmentation with Topic Models. JLCL, Vol. 27, No.1 (2012) 47-69
9. Zsibrita, J., Vincze, V., Farkas, R.: magyarlanc: A toolkit for morphological and dependency parsing of Hungarian. In: Proceedings of RANLP (2013) 763-771